T cell quantification from DNA sequencing predicts immunotherapy response

Robert Bentham[1,2*], Kevin Litchfield[2,3*], Thomas B K Watkins[4*], Emilia L Lim[2,4], Rachel Rosenthal[4], Carlos Martínez-Ruiz[1,2], Crispin T Hiley[2,4], Maise Al Bakir[4], Roberto Salgado[5,6], David A Moore[2,7,8], Mariam Jamal-Hanjani[2,8,9], TRACERx Consortium[10], Charles Swanton[2,4,8] and Nicholas McGranahan[1,2+]

1. Cancer Genome Evolution Research Group, Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, Paul O'Gorman Building, 72 Huntley Street, London, WC1E 6BT, UK.

2. Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, Paul O'Gorman Building, 72 Huntley Street, London, WC1E 6BT, UK.

3. The Tumour Immunogenomics and Immunosurveillance Lab, Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, Paul O'Gorman Building, 72 Huntley Street, London, WC1E 6BT, UK.

4. Cancer Evolution and Genome Instability Laboratory, The Francis Crick Institute, 1 Midland Rd, London NW1 1AT, UK

5. Department of Pathology, GZA-ZNA, Antwerp, Belgium

6. Division of Research, Peter MacCallum Cancer Centre, University of Melbourne, Melbourne, Victoria, Australia

7. Department of Cellular Pathology, University College London Hospitals, London, UK

8. Department of Medical Oncology, University College London Hospitals, 235 Euston Rd, Fitzrovia, London, United Kingdom, NW1 2BU, UK

9. Cancer Metastasis Lab, University College London Cancer Institute, Paul O'Gorman Building, 72 Huntley Street, London, WC1E 6BT, UK

10. A list of authors and their affiliations appears at the end of the paper

* These authors contributed equally

+ To whom correspondence should be addressed: nicholas.mcgranahan.10@ucl.ac.uk

Abstract

The immune microenvironment influences tumour evolution and can be both prognostic and predict response to immunotherapy[1,2]. However, measuring tumour infiltrating lymphocytes (TILs) is restricted by lack of appropriate data. Whole exome sequencing (WES) of DNA is frequently performed to calculate tumour mutational burden and identify actionable mutations. Here we develop a method for T cell fraction estimation from WES samples, utilising a signal from T cell receptor excision circle (TRECs) loss during VDJ recombination of the T cell receptor alpha (*TCRA*) gene. This score significantly correlates with orthogonal TIL estimates and is agnostic to sample type. Blood *TCRA* T cell fraction is higher in females and correlates with both tumour immune infiltrate and presence of bacterial sequencing reads. Tumour *TCRA* T cell fraction is prognostic in lung adenocarcinoma and using a meta-analysis of immunotherapy-treated tumours, we show that this score predicts immunotherapy response, providing value beyond tumour mutational burden. Applying this score to a multi-sample pan-cancer cohort revealed high diversity in immune infiltration within tumours. Subclonal loss of 12q24.31-32, encompassing *SPPL3*, was associated with reduced *TCRA* T cell fraction. Our method, T cell ExTRECT (T cell Exome TREC Tool), quantifies the T cell infiltrate of WES samples.

Introduction

Checkpoint inhibitors (CPIs) have emerged as revolutionary cancer treatments, acting to release the brakes on the immune system[3,4]. Clinical response, however, is not universal[5] and is principally governed by the presence of an immune stimulus, such as neoantigens, and an immune response, mediated by T cells[2]. While neoantigens can be predicted from WES[1], Until now, T cell quantification has required additional biological material, time, and expertise, adding to the cost of immunotherapy.

Here we propose a method for the estimation of the T cell fraction present in a WES sample. This method utilises a somatic copy number-based signal from VDJ recombination and the loss of TRECs. We explore the underlying features which predict T cell infiltrate in tumours and blood and evaluate determinants of immune heterogeneity within tumours. Finally, we demonstrate that our estimated T cell fraction can be used as a predictor of clinical response to CPI therapy.

Results

Inferring T cell fraction from WES data

T cell diversity, which is required for immune system recognition of foreign antigens, is a product of VDJ recombination, where segments within the T cell receptor genes recombine. The alpha chain of the T cell receptor is encoded by the *TCRA* gene (also known as *TRA*) and the result of VDJ recombination is the excision of unselected gene segments from *TCRA* as TRECs, with the T cell undergoing a deletion event within *TCRA*.

Tools to infer cancer somatic copy number alteration (SCNA)[6–9] rely on the read depth ratio (RDR), reflecting the log of the ratio of reads between the tumour sample and its matched control (e.g. buffy coat in a centrifuged blood sample). Deviation in the RDR from zero is assumed to reflect a tumour SCNA. However, within *TCRA* this assumption does not hold; a deviance in the RDR may reflect T cell specific deletion events and SCNA tools may thus erroneously infer tumour SCNA. Indeed, in the TRACERx100 cohort multiple SCNA within *TCRA* were inferred in 165/327 tumour regions (Extended Data Fig. 1a). The RDR deviated the most within segments frequently included within TRECs (Extended Data Fig. 1b-c). This suggests that most detected SCNAs within *TCRA* reflect a signal of relative T cell content rather than cancer SCNAs.

To exploit this signal to quantify T cell content we developed T cell ExTRECT (T cell Exome TREC Tool). T cell ExTRECT uses a modified RDR within *TCRA* to directly quantify T cell infiltrate in WES samples (Figure 1a), referred to as the *TCRA* T cell fraction. Unlike RNA-seq scores, the *TCRA* T cell fraction represents a direct quantification of the proportion of T cells within a sample. We identified no systematic significant difference in *TCRA* T cell fraction dependent on whether samples were fresh frozen or formalin-fixed paraffin-embedded (FFPE) (Methods, Extended Data Fig. 1d-e). T cell ExTRECT can be applied to any WES sample, thus permitting analysis of T cell fraction in both tumour and blood samples.

Validation of TCRA T cell fraction

To evaluate the accuracy of T cell ExTRECT, we used five orthogonal approaches.

First, to assess the ability to accurately determine the presence or absence of T cells within a sample, we used WES data from cell lines originating from T cell lymphoma (JURKAT, PEER, and HPB-ALL) and 14 colorectal cancer cell lines derived from HCT116 with varying degrees of genomic complexity[10,11]. All HCT116 cell lines had a calculated fraction of 0. Conversely, the three T cell lymphoma-derived cell lines had scores close to 1 (~0.95-0.96) (Extended Data Fig. 1f).

Second, we used an alternative DNA based method of inferring immune content[12], based on the number of reads that align to the CDR3 region following VDJ recombination (CDR3 VDJ score, Methods). In the TRACERx100[13] cohort (Extended Data Fig. 1g) we observed a significant positive correlation between *TCRA* T cell fraction and the CDR3 VDJ score (Extended Data Fig. 1h, ρ = 0.36, P = 1.4e-13). However, the CDR3 VDJ score was constrained by sequencing depth; the number of reads aligning to the CDR3 region was typically very low (1st quartile = 0, medium = 2, mean = 2.335, 3rd quartile = 3, maximum = 14).

Third, we simulated NGS data with a range of T cell fractions (Extended Data Fig. 2a-d). We observed a highly significant relationship between simulated and calculated T cell fraction (ρ = 0.99986, P < 2.2e-16, Extended Data Fig. 2b). Using downsampling and simulations, we found that the *TCRA* T cell fraction estimates remained consistent at coverage above and including 30X (ρ = 0.84, P = 1.4e-14) (Extended Data Fig. 2e-f). In contrast, the results from the CDR3 method were heavily skewed by sequencing coverage; when selecting the five samples with the highest CDR3 coverage and downsampling to 50X, only one sample with ≥ 3 CDR3 reads was detected (Extended Data Fig. 2g).

Fourth, to further confirm the accuracy of the *TCRA* T cell fraction, we evaluated its association with histopathology-derived TIL scores from H&E slides. Selecting the subset of tumour regions with both RNA-seq data and histopathology-derived TIL scores (147 regions), we evaluated how the *TCRA* T cell fraction, CDR3 VDJ score, and six RNA-seq based immune measures for CD8+ cells (Danaher[14], Davoli[15], xCell[16], TIMER[17], CIBERSORT[18], and EPIC[19]) compared to histopathology-derived TIL scores (Figure 1b). The Danaher CD8+ score had the strongest association (ρ = 0.49), followed by the *TCRA* T cell fraction (ρ = 0.41), Davoli (ρ = 0.4), xCell (ρ = 0.36), CIBERSORT (ρ = 0.23), TIMER (ρ = 0.2), CDR3 VDJ score (ρ = 0.2), and EPIC (ρ = 0.082).

Finally, the *TCRA* T cell fraction from WES was directly compared with RNA-seq methods and was found to have a significant positive relationship with multiple immune scores[1,14–19] with the strongest associations being with T cell related scores (Figure 1c).

Determinants of T cell content in blood

We next explored the key determinants of T cell immune infiltrate in matched control blood WES samples.

Within the TRACERx100[13] cohort, blood *TCRA* T cell fraction was significantly higher in females than males (Figure 2a, P = 0.0057, ES = 0.28) and we observed a trend for higher blood T cell fraction in LUSC compared to LUAD patients (Extended Data Fig. 3a, P = 0.066, ES = 0.19). We also observed a significant positive relationship between blood *TCRA* T cell fraction and matched tumour *TCRA* T cell fraction (Figure 2a, ρ = 0.42, P = 1.7e-05). These data suggest that tumour immune infiltrate may influence T cell levels in circulating blood or vice versa. We observed broadly consistent results in LUAD and LUSC TCGA[20,21] patients (Extended Data Fig. 3b-c).

To further examine the determinants of blood T cell fraction, we explored WES samples derived from both blood and physiologically normal oesophagus epithelia (PNE) tissue[22]. While blood samples exhibited a wide range of *TCRA* T cell fraction levels, the majority of PNE tissue had no detectable T cell infiltration (Extended Data Fig. 3d-e). Dividing the PNE samples by presence of T cell infiltration revealed a significant association with blood *TCRA* T cell fraction (Figure 2b, P = 0.021, ES = 0.29). Therefore, similarly to tumour samples, high levels of T cell infiltration in normal tissue may influence the TCRA T cell fraction observed in blood. In a linear model predicting T cell fraction in blood, only the infiltration level in normal tissue was significant (Extended Data Fig. 3f); no genomic factors, such as mutation burden or driver mutation status were predictive of T cell infiltration in PNE tissue (Extended Data Fig. 3g).

Viral or bacterial infections could also influence T cell levels in blood. To explore this we obtained normalised estimates for the abundance of microbial reads from blood and tumour samples from the LUAD and LUSC TCGA cohorts[23]. Blood samples with elevated microbial reads (> median, 6.81) had significantly higher blood TCRA T cell levels (Figure 2c, P =

0.00092, ES = 0.31, Wilcoxon test). No corresponding association was identified in tumour samples (Extended Data Fig. 3h, P = NS). No specific virus or bacteria were associated with blood *TCRA* T cell fraction. In tumour samples significant associations for bacteria of the genus *WIlliamsia* in LUAD (ρ = -0.17, P = 0.00011, FDR P = 0.013) and *Paeniclostridium* in LUSC (ρ = -0.2, P = 0.00013, FDR P = 0.015) were observed (Extended Data Fig. 3i-k). Both had higher normalised log-cpm values when *TCRA* T cell fraction was lower, suggesting they may be opportunistic species exploiting an immune-cold tumour microenvironment.

Determinants of tumour T cell content

Next, we explored factors influencing T cell infiltrate in tumour tissue. We utilised a recently published pan-cancer cohort of multi-sample data[24] to investigate both the extent and possible genomic basis for immune infiltrate heterogeneity. In total, we evaluated T cell infiltrate in 731 tumour samples from 178 tumours, from 12 cancer types (Extended Data Fig. 4a-b).

We classified each multi-sample tumour as uniformly hot (all samples ≥ 0.11, the mean TCRA T cell fraction), uniformly cold (all samples < 0.11) or heterogeneous. There was a significant difference in the proportion of these categories by cancer type (Figure 2d, chi-squared test: P = 1.62e-07) with ER+ breast cancer (BRCA ER+) tumours being the most heterogeneous (83%) and LUSC tumours being the least (22%). Clear differences in the prevalence and heterogeneity of immune infiltrate was observed across cancer types; for instance, while bladder cancer (BLCA) and LUAD had similar numbers of heterogeneous tumours (36% vs 37%), ~64% of BLCA tumours were uniformly immune-hot and 0% were uniformly immune-cold, whereas in LUAD 37% tumours were uniformly immune-cold and 25% uniformly immune-hot. This suggests that

for certain cancer types there is a highly localised immune infiltrate, which can be subject to considerable sampling bias.

Next, we examined the relationship between SCNAs and immune diversity. We restricted the analysis to tumours with at least three samples and a heterogeneous mixture of T cell infiltrate. Pairwise SCNA heterogeneity between any two samples was calculated as the sum of the proportion of the genome with unique SCNAs in either region. Pairs of tumour samples with a large disparity in *TCRA* T cell fraction (≥ the mean of all pairwise distances, 0.065) were associated with a larger differences in SCNA heterogeneity compared to matched region pairs with low TCRA T cell fraction heterogeneity (Figure 2e, All events: P = 0.0025, ES = 0.347; gain events: P = 0.0056, ES = 0.318; loss or LOH events: P = 0.028, ES = 0.253, n = 76).

To explore whether any specific subclonal SCNA were associated with immune depletion or activation, we identified cytobands that were subclonally lost or gained > 30 tumours in the pan-cancer multi-sample cohort (Extended Data Fig. 4c) and investigated whether specific SCNAs were associated with changes in *TCRA* T cell fraction. Subclonal loss of 12q24.31-32 was found to be significantly associated with decreased *TCRA* T cell fraction (Figure 2f: P = 5.9e-06, ES = 0.75).

RNA-seq analysis of the TRACERx100 cohort identified *SPPL3* as exhibiting the most significant differential expression between samples with and without subclonal 12q24.31-32 loss (Extended Data Fig. 4d). The absence of *SPPL3* has been found to augment B3GNT5 enzyme activity which upregulates cell surface glycosphingolipids that in turn impede class I HLA function and diminish CD8+ T cell activation[25]. Thus, these data suggest that subclonal loss of 12q24.31, encompassing *SPPL3*, may be selected in tumour evolution across cancer types (occuring in 18.7% of tumours within the cohort) as a mechanism of immune evasion.

T cell fraction is prognostic in LUAD

To explore the clinical utility of T cell ExTRECT, we considered whether the *TCRA* T cell fraction was prognostic in the TRACERx100 non-small cell lung cancer (NSCLC) cohort[13]. We categorised tumour regions as either 'hot' or 'cold' depending on whether *TCRA* T cell fraction was ≥ the mean in the cohort (0.081). In LUAD, we observed that patients harbouring an elevated number of immune-cold tumour regions were associated with significantly inferior prognosis (Figure 3, LUAD: ≥ 2 immune-cold regions, HR = 3.1, P = 0.0063 log-rank test, LUAD: ≥ 3 immune-cold regions HR = 7.3, P = 0.00024 log-rank test). In contrast, in LUSC patients there was no significant difference in survival. Using the median (0.074) as a threshold for immune hot or cold regions yielded similar results (Extended Data Fig. 5a). These results are consistent with previous analysis based on TIL scores inferred from computational pathology on the TRACERx100 cohort[26]. An association between high TCRA T cell fraction and good outcome was also observed in the TCGA LUAD (Extended Data Fig. 5b overall survival (OS): HR = 0.61, P = 0.0043, progression free survival (PFS): HR = 0.67 P = 0.016), but not LUSC cohort (Extended Data Fig. 5c). A range of possible thresholds yielded similar results (Extended Data Fig. 5d).

Consistent with the importance of the tumour region with the lowest immune infiltrate[26], the minimum, but not the maximum or mean, *TCRA* fraction across tumour regions was prognostic in the TRACERx100 cohort. Other continuous measures such as a *TCRA* T cell fraction divergence between tumour region score (Extended Data Fig. 5d, LUAD: HR = 2.2 P = 0.023 log-rank test) and a model combining both the minimum and maximum scores (Extended Data Fig. 5e, LUAD and LUSC: minimum HR = 0.5, P = 0.005, maximum HR = 1.5 P = 0.061; LUAD: minimum HR = 0.36, P = 0.016, maximum HR = 2.52, P = 0.029) reached significance,

suggesting that there is added predictive potential when considering the heterogeneity of the *TCRA* T cell fraction.

T cell fraction and response to CPIs

To further explore the clinical utility of T cell ExTRECT, we evaluated its ability to predict clinical response to CPIs. The CPI1000+ cohort[2] consists of 1070 CPI-treated tumours receiving either anti-CTLA-4, anti-PD-L1 or anti-PD-1 therapy across eight main cancer types (Extended Data Fig. 6a-b). A responder was defined as a patient with complete response (CR) or partial response (PR), while a non-responder was defined as stable disease (SD) or progressive disease (PD), on imaging by RECIST criteria[27].

Consistent with the importance of T cells in influencing response to CPIs, we observed a significantly higher (Figure 4a, P = 2.3e-07, ES = 0.17) tumour *TCRA* T cell fraction in responders. Likewise, immune-cold tumours (tumours with *TCRA* T cell fraction < 0.067, the mean *TCRA* T cell fraction), were significantly enriched for non-responders (Figure 4b, Fisher's exact test, odds ratio (OR) = 2.12, P = 2.25e-06).

Separating the cohort by the medians for both clonal TMB and *TCRA* T cell fraction revealed that the association between *TCRA* T cell fraction and clinical response was independent of clonal TMB (Figure 4b).

To evaluate the utility of T cell ExTRECT in comparison to RNA-seq based measurements, all studies with ≥ 10 samples from a cancer type with both RNA-seq and *TCRA* T cell fractions were selected for univariate meta-analyses (Figure 4c: 557 patients across 7 studies and 5 cancer types). *TCRA* T cell fraction (OR = 1.39, P = 0.00858), clonal TMB (OR = 1.59, P = 6.021e-05) and *CD8A* expression (OR = 1.45, P = 0.0004479) were all significantly associated with response.

To assess whether tumour *TCRA* T cell fractions improves prediction of response beyond clonal TMB and to a greater extent than *CD8A* expression we evaluated different linear models (Extended Data Fig. 6c). Only the clonal TMB + *TCRA* model was significant compared to clonal TMB alone (ROC test, P = 0.0028, GLM: clonal TMB + *TCRA*, AUC = 0.68, GLM: clonal TMB, AUC = 0.62). When examining the significance of the variables in all models, *TCRA* T cell fraction was more significant than *CD8A* (GLM: clonal TMB + *TCRA*, P = 4.62e-05; GLM: clonal TMB + *CD8A*, P = 0.000431) and when combined into a multivariable model, TCRA T cell fraction remained significant, but *CD8A* expression did not (*TCRA*, P = 0.00601, *CD8A*, P = 0.06246).

Finally, we assessed the predictive potential of the *TCRA* T cell fraction in a combined NSCLC CPI cohort (Extended Data Fig. 6d-e) lacking any RNA-seq immune measures. In univariate analyses, (Figure 4d), *TCRA* T cell fraction (OR = 1.44, P = 0.0071) and blood *TCRA* T cell fraction (OR = 1.39, P = 0.015) were significantly associated with response to CPI. Tumour *TCRA* T cell fraction had OR > 1 in two of three cohorts while blood *TCRA* T cell fraction had OR > 1 in all three cohorts.

Taken together, these results suggest the *TCRA* T cell fraction can be used as a substitute for RNA-seq measures of CD8+ infiltrate, and, moreover, *TCRA* T cell fraction estimation adds prognostic value to TMB estimates.

Discussion

In summary, we present a method, T cell ExTRECT, by which DNA WES can be harnessed to study the immune microenvironment. T cell ExTRECT provides an accurate estimate of immune infiltrate which shows clinical utility. We find tumour *TCRA* T cell fraction is prognostic in LUAD and validate this finding in the TCGA LUAD cohort. Relatedly, we find the *TCRA* T cell fraction is associated with response to CPI in a pan-cancer cohort and improves upon the

predictive value of clonal TMB. T cell ExTRECT enables the T cell fraction to be calculated in any WES sample. Leveraging this, we demonstrate that T cell fraction in blood is heterogeneous, associated with microbial infections and was found to be significantly higher in females than males in TRACERx100 NSCLC patient data, consistent with previous findings[30,31]. Our analysis of blood samples in the lung CPI cohort revealed that blood *TCRA* T cell fraction is predictive of response to immunotherapy.

The T cell ExTRECT method has limitations. While the tool provides a quantification of the proportion of T cells in a sample, it cannot distinguish neoantigen-reactive from bystander T cells, and is unable to detect clonotypes. Further, T cell ExTRECT loses fidelity below 30X sequencing depth. Nevertheless, this relatively low depth means it should be applicable to most DNA sequencing datasets. T cell ExTRECT has so far only been optimised for WES, but further work will extend the method to whole-genome and to other species including much studied model organisms. T cell ExTRECT has clear applications in the immuno-oncological exploration of tumour samples, however it could also be utilised in a wider clinical setting, such as newborn screening of severe combined immunodeficiency disease[32].

In summary, our approach, T cell ExTRECT, could have important applications in both basic and translational research by providing a cost-effective technique to characterise immune infiltrate alongside somatic changes, without the need for RNA sequencing.

Methods

A detailed and full description of the T cell ExTRECT method is given in Supplementary Information.

Statistics

All statistical tests were performed in R 3.6.1. No statistical methods were used to predetermine sample size. Tests involving correlations were done using 'stat_cor' from R package ggpubr (v0.4.0) with the Spearman's method. Tests involving comparisons of distributions were done using 'stat_compare_means' using either 'wilcox.test' using the unpaired option, unless otherwise stated. Effect sizes for the corresponding Wilcoxon tests were measured using the 'wilcox_effsize' function from the rstatix package (v0.6.0). Hazard ratios and P values were calculated with the 'survival' package (v3.2-3) for both the Kaplan-Meier curves and Cox proportional hazard model. For all statistical tests, the number of data points included are plotted or annotated in the corresponding figure. Plotting and analysis in R also made use of the ggplot2 (v3.3.3), dplyr (v1.0.4), tidyr (v1.1.1), gridExtra (v2.3) and gtable (v0.3.0) packages.

Fresh frozen vs FFPE samples

To test that the *TCRA* T cell fraction was reliable and consistent for both fresh frozen and FFPE samples the non-GC corrected *TCRA* T cell fractions were calculated for six different studies within the CPI1000+ cohort. Three of these studies utilised WES derived from FFPE tissues (n = 460) while the other three utilised WES samples derived from fresh frozen tissue (n = 357).

Fitting a linear model to predict *TCRA* T cell fraction by histology and FFPE status (Extended Data Fig. 1i) revealed that cancer type however was the main driver of this significance with FFPE status not being significant. Additionally, for melanoma and bladder tumours that had FFPE and fresh frozen WES samples there was no significant difference found (Extended Data Fig. 1f). This led us to conclude that whether a WES sample is derived from fresh frozen or FFPE tissue does not significantly affect the values of the *TCRA* T cell fraction calculated by T cell ExTRECT.

Calculation of CDR3 VDJ scores

The procedure outlined in Levy et al.[12] was followed to calculate the CDR3 VDJ scores. First reads aligning to *TCRB* (hg19:chr7:142000817-142510993) and unaligned reads were extracted with samtools, this resulting bam was converted to fastq using bedtools and then the tool IMSEQ (v1.1.0)[33] was used on the resulting output to identify VDJ recombinant reads aligning to the CDR3 region, the number of aligned reads was than normalised by the total number of reads in the original bam file (as measured by samtools flagstat) to create the CDR3 VDJ scores.

Kraken TCGA analysis

Pre-processed microbiome data output from the Kraken[34] analysis performed by Poore et al.[23] was downloaded from ftp://ftp.microbio.me/pub/cancer_microbiome_analysis/.

To create the high and low Kraken microbiome groups for both the blood and tumour samples the file Kraken-TCGA-Voom-SNM-Most-Stringent-Filtering-Data.csv was downloaded containing normalised log-cpm values, for each sample the rows were summed giving a overall 'microbiome' score. The samples were then divided into high and low groups based on the median of this score.

To investigate the role of any individual microbial species in influencing TCRA T cell fraction a reduced list of the species from the Kraken-TCGA-Voom-SNM-Most-Stringent-Filtering-Data.csv file were selected, by removing all species with less than 1000 total raw reads in the TCGA LUAD and LUSC cohort as called from the raw data file Kraken-TCGA-Raw-Data-17625-Samples.csv. This left a total of 59 microbial species that were individually tested for association with *TCRA* T cell fraction using Spearman's correlation for both LUAD and LUSC blood and tumour samples.

TRACERx100 patients

The first 100 patients prospectively analysed by the NSCLC TRACERx study (https://clinicaltrials.gov/ct2/show/NCT01888601, approved by an independent research ethics committee, 13/LO/1546) were used in this study. This is identical to the 100 patient cohort originally described in Jamal-Hanjani et al[13].

Describing this cohort in brief, informed consent was a mandatory requirement for entry into the TRACERx study. This NSCLC cohort consisted of 68 males and 32 female patients with a median age of 68. Finally, the cohort is predominantly made up of early-stage tumours (Ia (26), Ib (36), IIa (13), IIb (11), IIIa (13) and IIIb (1)) and 28 patients also had adjuvant therapy.

TRACERx100 WES and RNA-seq samples

Both WES (aligned to hg19) and RNA-seq samples were obtained from the TRACERx study for the first 100 patients, the method for processing these samples is as previously described[13]. Notably for the WES samples, exome capture was performed using a custom version of Agilent Human All Exome V5 kit as per the manufacturer instructions.

TCGA LUAD and LUSC cohorts

Aligned BAM files (hg38) from the TCGA LUAD and LUSC cohorts were downloaded from the genomic data commons (dataset ID: phs000178.v10.p8). Sample purity and ploidy calls were generated from ASCAT (v2.4.2) from a previous analysis of the TCGA data[35], in short Affymetrix SNP6 profiles from paired tumour-normal samples (dataset ID: phs000178.v10.p8) were processed by PennCNV libraries[36] to obtain BAFs and log ratios which were GC corrected before being processed with ASCAT[6].

Cancer cell line data

The non-T cell derived colorectal cancer cell lines HCT116 were sequenced with Illumina HiSeq 2500 and aligned with bwa mem using hg19 as described in López et al.[10]. The T cell derived cell lines were from the dataset were described in Ghandi et al.[11] and downloaded

from the Sequence Read Archive (SRA) under accession number PRJNA523380. Cell lines derived from T cells were chosen ensuring that any cell line derived from precursor T cell acute lymphoblastic leukemia were excluded as these have not undergone VDJ recombination. This process led to WES data from three cell lines being chosen: JURKAT, HPB-ALL, and PEER.

Due to the difficulty of running ASCAT without matching germline samples, the naïve *TCRA* T cell fraction was used for all cell line work.

Multi-sample tumour cohort of patients

The multi-sample pan-cancer cohort (Extended data Fig. 4b) was created by combining the TRACERx cohort with a subset of the cohort presented recently by Watkins et al.[24]. Tumours were included if they had at least two regions sequenced in the primary tumour for which it was possible to calculate the *TCRA* T cell fraction using T cell ExTRECT. The final cohort therefore consisted of a multi-region primary tumour data set with the addition of any metastasis samples that were also sequenced for these patients.

Besides TRACERx100 the following datasets were combined into the final multi-sample pan-cancer cohort:

1. Brastianos et al.[37] - a cohort focused on studying brain metastasis originating from different histologies, only tumours with multi-region primary samples from this cohort were included.
2. Gerlinger et al.[38,39] - A multi-sample primary cohort of renal clear cell carcinoma (KIRC) patients.
3. Harbst et al.[40] - A multi-region primary cohort of skin cutaneous melanoma (SKCM) patients.
4. Lamy et al.[41] - A multi-region primary cohort of bladder cancer patients (BLCA)

5. Savas et al.[42] - A multi-sample cohort of ER+ and triple-negative breast cancer patients (BRCA ER+ and TNBC)
6. Suzuki et al.[43] - A multi-region primary cohort of glioma.
7. Turajlic et al[44]. - A multi-region primary cohort of clear cell renal cell carcinoma (KIRC).
8. Messaoudene et al.[45] - A multi-region primary cohort of HER2+ and ER+ breast cancer patients.

Selection of subregions for multi-region sequencing in different data sets

In all of the multi-region cohorts regions were selected though by different methods (see associated publications) with two main criteria in mind, first that tumour content be maximised at the expense of stromal in order to assure good quality mutation and copy number analysis for the main goal of the genomic analysis and second that each region represent a physically separate and distinct part of the tumour. In cases where these were not at separate sites different measures were used. In the TRACERx100 cohort for example regions sequenced were a minimum of 3mm apart.

Identification of gain, loss, and LOH events in a pan-cancer multi-sample cohort

Analysis of whole-exome sequencing was performed as described previously[13]. Copy-number segmentation, tumour purity and ploidy for each sample were estimated using ASCAT[6] as described previously[13]. These data were used as input to a multi-sample SCNA estimation approach to produce genome-wide estimates of the presence of loss of heterozygosity as well as loss, neutral, gain and amplification copy-number states relative to sample ploidy. The log ratio values present in each copy-number segment with ≥5 log ratio values in all samples of a tumour were examined relative to three sample-ploidy-adjusted log ratio thresholds using one-

tailed t-tests with a P < 0.01 threshold. These log ratio thresholds were equivalent to <log2[1.5/2] for losses, >log2[2.5/2] for gains in a diploid tumour. Any segment not classified as a loss or gain were classed as neutral. For each segment, these relative to ploidy definitions were combined with loss of heterozygosity detection across all samples from a single tumour.

Pairwise subclonal SCNA scores

To calculate pairwise subclonal SCNA measures, the classifications outlined in the previous methods section were used to create three groups of pairwise subclonal SCNA scores. First, we considered any segment affected by any of gain or loss relative to ploidy or LOH as aberrant and compared each pair of regions from a single patient's disease, classifying aberrant areas as clonal if aberrant in both samples or subclonal if aberrant in only one sample. This same process was repeated for gains relative to ploidy alone and then losses relative to ploidy and LOH considered together.

Cytoband-level SCNA analysis

To enable comparisons across tumours, segments were mapped to hg19 cytobands. If multiple segments mapped to a cytoband, the SCNA status (gain or loss relative to ploidy) of the segment with the largest overlap with the cytoband was chosen.

For the SCNA gain and loss analysis, cytoband level events were selected if they occurred subclonally across the entire cohort greater than 30 times. Bands passing this threshold within the same region (e.g. all cytobands on 1p36) were then grouped together. A Wilcoxon paired test was used to assess whether the tumour regions within a single patient with the subclonal SCNA events had a significant difference in *TCRA* T cell fraction to those regions without the event.

Selection of multi-sample tumours with heterogeneous immune infiltration

To be included a tumour had to have at least 3 regions sequenced and meet the following two requirements, 1) have a pair of regions with a large change in immune infiltration as defined as having ≥ 0.065 difference in *TCRA* T cell fraction, and 2) have a pair of regions with a small or no change in immune infiltration as defined as having < 0.065 difference in *TCRA* T cell fraction. An example of a tumour matching this requirement would be one with three regions R1, R2 and R3 with *TCRA* T cell fractions of 0.01, 0.01 and 0.2 respectively. The R1-R2 pair has a difference in *TCRA* T cell fraction of 0 while the R1-R3 and R2-R3 pairs would both have a large difference of 0.19. Within the multi-sample tumour cohort 76 patients matched these criteria.

RNA-seq differential gene expression analysis for patients with subclonal 12q24.31-32 loss

Differential gene expression analysis was performed on the TRACERx100 RNA-seq patients with subclonal 12q24.31-32 loss. Using R 4.0.0, first the edgeR R package (version 3.32.1) was used for sample-specific TMM normalisation, any genes with low expression were then filtered out using the standard edgeR filtering method before using the Limma-Voom method from the limma R package (version 3.46.0) to calculate the Voom fit and obtain p-values for the gene expression differences. The comparison controlled for patient and histology as blocking factors and p-values were FDR corrected for multiple testing. Results were then visualised with the R EnhancedVolcano package (version 1.8.0).

CPI1000+ meta-analysis of cohorts

The CPI1000+ cohort is fully described in Litchfield et al.[2] and contains the following datasets:

1. Snyder et al.[46], an advanced melanoma anti-CTLA-4 treated cohort.
2. Van Allen et al.[47], an advanced melanoma anti-CTLA-4 treated cohort.
3. Hugo et al.[48], an advanced melanoma anti-PD-1 treated cohort.

4. Riaz et al.[49], an advanced melanoma anti-PD-1 treated cohort.
5. Cristescu et al.[50], an advanced melanoma anti-PD-1 treated cohort.
6. Cristescu et al.[50], an advanced head and neck cancer anti-PD-1 treated cohort.
7. Cristescu et al.[50] “all other tumour types” cohort (from KEYNOTE-028 and KEYNOTE-012 studies), treated with anti-PD-1.
8. Snyder et al.[51], a metastatic urothelial cancer anti-PD-L1 treated cohort.
9. Mariathasan et al.[52], a metastatic urothelial cancer anti-PD-L1 treated cohort.
10. McDermott et al.[53], a metastatic renal cell carcinoma anti-PD-L1 treated cohort.
11. Rizvi et al.[29], a non-small cell lung cancer anti-PD-1 treated cohort.
12. Hellman et al., a cohort of non-small cell lung cancer samples treated with anti-PD-1 used by Litchfield et al.[2].
13. Le et al.[54],a colorectal cancer cohort treated with anti-PD-1 therapy.

Of these studies Snyder et al.[51] was excluded from the analysis due to extremely poor coverage within the *TCRA* gene. Additionally, 55 patients were either on treatment at the time of the biopsy or had prior treatment with CPIs and were removed from the analysis. All samples were aligned to hg19 using bwa mem (v0.7.15) with purity and SCNA data calculated using ASCAT as described in Litchfield et al.[2].

Notably, 953/1070 (89%) samples had WES data, 888/1070 (83%) had sufficient purity and coverage to enable copy number calculation enabling the *TCRA* T cell fractions to be calculated. 643/1070 (60%) of these samples had matched RNA-seq data allowing orthogonal assessment of T cell estimates.

For an extension to this dataset, Shim et al.[28] a NSCLC anti-PD-1 treated cohort was added for a specific NSCLC analysis. In this entire cohort mutations were called as either clonal or subclonal using PyClone as described by Litchfield et al.[2].

Orthogonal immune measures

RNA-seq signatures

We used the method of Danaher et al.[12] as our primary method of estimating T cell content from RNA-seq measures as it has been previously demonstrated that this is most strongly correlated to TIL scores calculated in TRACERx[1]. Other RNA-seq signatures tested against the *TCRA* T cell fractions were theDavoli method[15], xCell[16], TIMER[17] and EPIC[19] and CIBERSORT[18].

Histopathology-derived TIL scores

TILs were estimated, as previously described in Rosenthal et al.[1], from histopathology slides using internationally established guidelines, developed by the International Immuno-Oncology Biomarker Working Group[55]. In brief, the relative proportion of stromal area to tumour area was determined from the pathology slide of a given tumour region. TILs were reported for the stromal compartment (= percent stromal TILs). The denominator used to determine the percent stromal TILs was the area of stromal tissue (that is, the area occupied by mononuclear inflammatory cells over total intratumoral stromal area) rather than the number of stromal cells (that is, the fraction of total stromal nuclei that represent mononuclear inflammatory cell nuclei). This method has been demonstrated to be reproducible among trained pathologists[56]. An inter-person concordance was performed, and this demonstrated high reproducibility. The International Immuno-Oncology Biomarker Working Group has developed a freely available training tool to train pathologists for optimal TIL assessment on haematoxylin–eosin slides (www.tilsincancer.org).

Univariate and multivariable model for CPI response

For the univariate model an adapted procedure from Litchfield et al.[2] was followed with the main difference being that only samples with complete data (RNA-seq for *CD8A*, clonal TMB and *TCRA* T cell fraction) were included. The univariate model meta-analysis was conducted using R package ‘meta’ (version 4.13-0). The multivariable model was created with general

linear models using the function `glm` from the 'stats' R package using default values. The R package 'ROCR' (version 1.0-11) was used for the ROC curve analysis.

Code

The code used to produce *TCRA* T cell fraction scores is available for academic non-commercial research purposes upon reasonable request.

All other code used in the analysis and to produce figures is available at: https://github.com/McGranahanLab/T-cell-ExTRECT-figure-code-2021

Data availability

The RNA-seq data and WES data (in each case from the TRACERx study) generated, used or analysed during this study are not publicly available and restrictions apply to the availability of these data. Such RNA-seq and WES data are available through the Cancer Research UK & University College London Cancer Trials Centre (ctc.tracerx@ucl.ac.uk) for academic non-commercial research purposes upon reasonable request, and subject to review of a project proposal that will be evaluated by a TRACERx data access committee, entering into an appropriate data access agreement and subject to any applicable ethical approvals.

Details of all other datasets obtained from third parties used in this study can be found in Extended Data Table 1. Clinical trial information (if applicable) is also available within the associated publications described in Extended Data Table 1.

Figure Legends

Figure 1 – Overview and validation of T cell ExTRECT

a, Overview of how VDJ recombination signal is identified from read depth within *TCRA* in T cell fraction calculation. **b,** Association with histopathology TIL scores and measures of CD8+

T cell content from either RNA-seq (Danaher, Davoli, EPIC, TIMER, CIBERSORT and xCell) or DNA (T cell ExTRECT and CDR3 VDJ score). **c,** Association between *TCRA* T cell fraction with RNA-based scores for immune cell types (Danaher[14], Davoli[15], EPIC[19], TIMER[17], CIBERSORT[18], and xCell[16]) ordering determined by strength of association (Spearman's Rho coefficient) with *TCRA* T cell fraction.

Figure 2: Determinants of T cell fraction.

a, TRACERx100 blood *TCRA* T cell fraction predictors. **b,** Association of *TCRA* T cell fraction in PNE with blood *TCRA* T cell fraction. **c,** Microbial reads from Kraken versus blood *TCRA* T cell fraction (n = 111). **d,** Proportion of tumours uniformly immune-hot, uniformly immune-cold or heterogeneous (Methods). **e,** Multi-sample tumours (n = 76) with heterogeneous immune infiltrate defined as having both a pair of regions with pairwise *TCRA* T cell fraction difference < 0.065 and another with pairwise difference ≥ 0.065, versus pairwise SCNA heterogeneity score (Methods). Threshold 0.065 being the mean of all pairwise differences between regions. **f,** *TCRA* T cell fraction difference between regions with or without subclonal loss of 12q24.31-32. All Wilcoxon tests two sided and boxplots represent lower quartile, median, and upper quartile.

Figure 3 – Prognostic value of *TCRA* T cell fraction within LUAD but not LUSC

TRACERx100 multi-region LUAD (top) and LUSC (bottom) Kaplan-Meier curves divided by the number of immune-cold regions in the tumour (increasing left to right). Immune-hot and immune-cold regions defined using threshold of the mean of all tumour regions (0.08095). Patients in Kaplan-Meier analyses were restricted to those with total regions greater than the number of immune-cold regions used in defining the threshold.

Figure 4 – *TCRA* T cell fraction is predictive of survival and response to immunotherapy

a, Violin plot showing the tumour *TCRA* T cell fraction for non-responders versus responders across the CPI1000+ cohort, dotted black line shows mean *TCRA* T cell fraction (0.067) **b,** Tumour *TCRA* T cell fraction versus clonal TMB, dashed lines divide cohort into four quadrants with high/low clonal TMB and immune-hot/immune-cold tumours separated by the median values. Inset pie charts indicate the percentage of patients demonstrating CPI response. **c,** Univariate meta-analysis of predictors of CPI response across multiple cohorts with ≥ 10 patients of a cancer type and both DNA and RNA-seq data. Left panel: forest plot of OR values from different clinical factors with associated p-values in terms of predictive value of response. Right panel: heatmap of OR values across individual studies from the CPI1000+ dataset, focusing on cohorts with both RNA-seq and *TCRA* T cell fraction. **d,** Univariate meta-analysis across three CPI lung datasets with DNA but no RNA- seq data.

Extended Data Fig. 1: Overview and validation of T cell ExTRECT

a, Outline of quantification of the *TCRA* T cell fraction utilising VDJ recombination and TRECs. *top:* Schematic demonstrating how RDR signals are used to detect SCNA gain or loss events in a standard tumour and matched control sample analysis. In this analysis cells consist of three distinct cell types: tumour cells, T cells and all other stromal cells. *bottom:* Schematic of how this same process works when focussing on the *TCRA* gene in relation to VDJ recombination and TRECs, the lower right panel indicates an increased number of breakpoints detected in the TRACERx100 dataset within the *TCRA* gene relative to surrounding areas of 14q, suggesting that the TREC signal is captured. **b, c,** Plots showing examples of RDR in two TRACERx100 regions demonstrating either increased levels of T cell content in blood compared to matched tumour (**b**) or increased levels of T cell content in tumour compared to matched blood (**c**). VDV segments refer to variable segments in both the TCRα and TCRδ locus. **d,** *TCRA* T cell fraction (non-GC corrected) value for FFPE and fresh frozen samples

for bladder and melanoma tumours within the CPI1000+ cohort (bladder: n = 228, melanoma: n= 297, two sided wilcoxon test used, boxplot shows lower quartile, median and upper quartile values). **e,** Summary of linear model for prediction of non-GC corrected *TCRA* T cell fraction from histology and FFPE sample status within the CPI cohort. **f,** Pie charts of calculated *TCRA* T cell fraction from WES of either T cell-derived cell lines or non-T cell derived cell lines, all HCT116 cell lines had calculated fractions < 1 e-15. **g,** Overview of samples in the TRACERx100 cohort. **e,** Association of the CDR3 VDJ read score based on the iDNA method to *TCRA* T cell fraction in TRACERx100, error bands represent the 95% confidence interval of the fitted linear model.

Extended Data Fig. 2: Accuracy of *TCRA* T cell fraction by copy number and depth

a, Simulated log RDR from a sample consisting of 24% T cells, 75% tumour, and 1% non-T cell stroma (*TCRA* copy number = 1). **b,** Calculated *TCRA* T cell fraction versus actual T cell fraction value for simulated data **c,** Difference between calculated naïve T cell fraction and actual fraction for range of tumour purities and local tumour copy number states at the *TCRA* locus. **d,** Difference between *TCRA* T cell fraction and actual fraction for a range of tumour copy number and purities. **e,.** Downsampling of 5 TRACERx100 regions to different depths. **f,** Downsampling of simulated data to different depth levels. **g,** Downsampling of the 5 TRACERx100 regions that with the highest CDR3 read counts to different depths and the resulting CDR3 read counts.

Extended Data Fig. 3: Extended analysis on determinants of *TCRA* T cell fraction

a, Association of blood *TCRA* T cell fraction to histology in TRACERx100 (n = 93 LUAD and LUSC patients, two sided wilcoxon test used for P value). **b,** Predictors of blood *TCRA* T cell fraction in TCGA LUAD and LUSC cohort (left panel: n = 1017, middle panel: n = 976, right panel: n = 714). **c,** Overview of samples in the TCGA LUAD and LUSC cohort. **d,** Summary of mean *TCRA* T cell fraction in PNE cohort. **e,** Overview plot of PNE cohort containing multi-region microdissected tissue paired with normal blood samples. **f,** Summary of linear model

for predicting blood *TCRA* T cell fraction, PNE infiltration defined as *TCRA* T cell fraction > 0.001, ESCC = Oesophageal squamous cell carcinoma, HGD = high grade dysplasia. **g,** Linear model for *TCRA* T cell fraction in PNE samples from genomic factors. **h,** Association of microbial reads from Kraken with *TCRA* T cell fraction in tumour samples (n = 880). **i,** -Log10 p-values for 59 microbial species tested for association with *TCRA* T cell fraction in blood and tumour sample in LUAD and LUSC. Red line represents the significance threshold at P = 0.000423. **j,** The significant hit *Willamsia* in LUAD tumours, red dots represent samples where reads were detected while blue represent samples with no reads detected (n = 501). **k,** The significant hit *Paeniclostridium* in LUSC tumours (n = 379). All wilcoxon tests two sided and boxplots represent lower quartile, median and upper quartile.

Extended Data Fig. 4: Subclonal SCNAs and T cell infiltration

a, Overview of immune heterogeneity across multi-sample pan-cancer cohort with tumour regions ranked by *TCRA* T cell fraction, *upper panel:* histogram of entire cohort, *lower panel:* tumour regions grouped by patients with solid horizontal lines joining regions from the same patient, each line includes 2 or more tumour region and dashed red line is at the mean *TCRA* T cell fraction in the cohort (0.11). **b,** Overview of patients in the multi-sample pan-cancer cohort. **c,** Lower panel: number of tumours in pan-cancer multi-sample cohort with subclonal gains (dark red) or losses (dark blue) across the genome, horizontal lines signify the regions which have more than 30 tumours (Methods) with subclonal gains or losses. *Upper panel*: -log10(p-value) of the 160 cytoband regions tested for association between *TCRA* T cell fraction and subclonal gains (dark red points) or losses (dark blue points). Red horizontal line marks significance threshold, only one region is significant, a loss event on chromosome 12q24.31-32. **d,** Volcano plot for the RNA-seq analysis in the TRACERx100 cohort between regions with 12q24.31-32 loss and regions without, genes within the locus are labeled, dotted lines at fold change of 0.25 and adjusted P = 0.05.

Extended Data Fig 5 : Association of *TCRA* T cell fraction with prognosis

a, Kaplan-Meier curves for the multi-region TRACERx100 cohort for LUAD (top) and LUSC (bottom) divided by the number of cold regions in the tumour. Hot and cold regions were defined by using the median of all the tumour regions (0.0736) as a threshold. In each Kaplan-Meier curve the included patients were restricted to those with total regions greater than the number of cold regions used in defining the threshold. **b,** Kaplan-Meier curves for overall and progression free survival in the TCGA LUAD cohort, dividing the cohort into immune hot and cold groups using the mean of the TCGA LUAD cohort (0.109) as a threshold. **c,** Kaplan-Meier curves for the TCGA LUSC, and TCGA LUAD & LUSC cohorts for overall and progression free survival using the mean of the TCGA LUAD cohort (0.109) as a threshold for distinguishing hot and cold tumours. **d,** Log2(Hazard ratios) from Kaplan-Meier plots for the TCGA separating the tumour samples into hot and cold based on different thresholds from 0 to 0.16 in steps of 0.0025 for overall and progression free survival. **e,** Hazard ratios of separate Cox regression models relating disease free survival to different multi-region measures related to the *TCRA* T cell fraction in the entire TRACERx100 cohort as well as the LUAD and LUSC patients separately. *TCRA* divergence score is defined as the maximum divided by the upper 95% confidence interval of the minimum. **f,** Hazard ratios of separate Cox regression models for *TCRA* T cell fraction for the TCGA LUAD and LUSC cohort for both overall survival (OS) and progression free survival (PFS).

Extended Data Fig 6: Overview of CPI1000+ cohort

a, Cohort overview of the CPI1000+ dataset. **b,** Overview of samples in the CPI1000+ cohort excluding Snyder et al., 2017 and those with prior CPI treatment. **c,** ROC plot of GLM models for predicting CPI response (blue: clonal TMB, red: clonal TMB + *TCRA* T cell fraction, green: clonal TMB + *CD8A* expression). **d,** Cohort overview of the CPI lung dataset, red lines in upper panel reflect the median *TCRA* T cell fraction in patients with (0.10) or without (0.0070) a response to CPI, note that Tumour *TCRA* T cell fraction particularly in non-responders is often zero. **e,** Overview of patients in the CPI Lung cohort.

Extended Data Table 1: Original source publications

Original source publications (excluding TRACERx studies) containing the sequencing data used in either the multi-sample pan-cancer cohort, PNE cohort or the CPI1000+ cohort. Studies including lung cancer patients used in the lung CPI cohort are noted.

Author contributions

R.B. helped conceive the study, designed and conducted the bioinformatic analysis, and wrote the manuscript. K.L. curated the CPI1000+ cohort used in the study and provided considerable bioinformatic support on its analysis. T.B.K.W. provided considerable bioinformatic support on the analysis of the multi-sample pan-cancer cohort and helped conceive the study and write the manuscript. T.B.K.W. and E.L.L jointly curated the multi-sample pan-cancer cohort used in the study. R.R. and C.M.-R. provided considerable bioinformatic support in the transcriptomic analysis performed in the study, providing RNA-seq immune score metrics and assisting with the RNA-seq gene expression analysis respectively. R.S., M.A.B., D.A.M., and C.T.H. jointly analysed histopathology-derived TIL estimates. C.S. helped provide study supervision and helped direct the avenues of bioinformatics analysis and also gave feedback on the manuscript. N.M conceived and supervised the study and helped write the manuscript.

Competing interests

D.A.M. reports speaker fees from AstraZeneca. M.A.B. has consulted for Achilles Therapeutics. R.R. has consulted for and has stock options in Achilles Therapeutics. K.L. reports speaker fees from Roche Tissue Diagnostics. C.T.H. has received speaker fees from AstraZeneca. M.J.-H. is a member of the Scientific Advisory Board and Steering Committee for Achilles Therapeutics. N.M. has stock options in and has consulted for Achilles Therapeutics and holds a European patent in determining HLA LOH (PCT/GB2018/052004).

C.S. acknowledges grant support from Pfizer, AstraZeneca, Bristol Myers Squibb, Roche-Ventana, Boehringer-Ingelheim, Archer Dx Inc. (collaboration in minimal residual disease sequencing technologies) and Ono Pharmaceutical; is an AstraZeneca Advisory Board Member and Chief Investigator for the MeRmaiD1 clinical trial; has consulted for Amgen, Pfizer, Novartis, GlaxoSmithKline, MSD, Bristol Myers Squibb, AstraZeneca, Illumina, Genentech, Roche-Ventana, GRAIL, Medicxi, Bicycle Therapeutics, Metabomed and the Sarah Cannon Research Institute; has stock options in Apogen Biotechnologies, Epic Bioscience and GRAIL; and has stock options and is co-founder of Achilles Therapeutics. C.S. holds patents relating to assay technology to detect tumour recurrence (PCT/GB2017/053289); to targeting neoantigens (PCT/EP2016/059401), identifying patent response to immune checkpoint blockade (PCT/EP2016/071471), determining HLA LOH (PCT/GB2018/052004), predicting survival rates of patients with cancer (PCT/GB2020/050221), to treating cancer by targeting Insertion/deletion mutations (PCT/GB2018/051893); identifying insertion/deletion mutation targets (PCT/GB2018/051892); methods for lung cancer detection (PCT/US2017/028013); and identifying responders to cancer treatment (PCT/GB2018/051912).

Acknowledgements

R.B. is supported by the NIHR BRC at University College London Hospitals. K.L. is funded by the UK Medical Research Council (MR/P014712/1), Rosetrees Trust and Cotswold Trust (A2437) and Cancer Research UK (C69256/A30194). T.B.K.W. was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001169), the UK Medical Research Council (FC001169) and the Wellcome Trust (FC001169) as well as the Marie Curie ITN Project PLOIDYNET (FP7-PEOPLE-2013, 607722), Breast Cancer Research Foundation (BCRF), Royal Society Research Professorships Enhancement Award (RP/EA/180007) and the Foulkes Foundation. E.L.L. receives funding from NovoNordisk Foundation (ID 16584). R.R. is supported by Royal Society Research Professorships

781 Enhancement Award (RP/EA/180007). C.M.R is supported by Rosetrees. C.T.H. is supported
782 by the NIHR BRC at University College London Hospitals. M.J.-H. has received funding from
783 Cancer Research UK, National Institute for Health Research, Rosetrees Trust, UKI NETs and
784 NIHR University College London Hospitals Biomedical Research Centre. N.M. is a Sir Henry
785 Dale Fellow, jointly funded by the Wellcome Trust and the Royal Society (Grant Number
786 211179/Z/18/Z) and also receives funding from Cancer Research UK, Rosetrees and the
787 NIHR BRC at University College London Hospitals and the CRUK University College London
788 Experimental Cancer Medicine Centre. C.S. is a Royal Society Napier Research Professor.
789 His work was supported by the Francis Crick Institute, which receives its core funding from
790 Cancer Research UK (FC001169), the UK Medical Research Council (FC001169) and the
791 Wellcome Trust (FC001169). C.S. is funded by Cancer Research UK (TRACERx, PEACE and
792 CRUK Cancer Immunotherapy Catalyst Network), Cancer Research UK Lung Cancer Centre
793 of Excellence, the Rosetrees Trust, Butterfield and Stoneygate Trusts, NovoNordisk
794 Foundation (ID16584), Royal Society Research Professorships Enhancement Award
795 (RP/EA/180007), the NIHR BRC at University College London Hospitals, the CRUK-UCL
796 Centre, Experimental Cancer Medicine Centre and the Breast Cancer Research Foundation
797 (BCRF). This research is supported by a Stand Up To Cancer-LUNGevity-American Lung
798 Association Lung Cancer Interception Dream Team Translational Research Grant (SU2C-
799 AACR-DT23-17). Stand Up To Cancer is a program of the Entertainment Industry Foundation.
800 Research grants are administered by the American Association for Cancer Research, the
801 Scientific Partner of SU2C. C.S. also receives funding from the European Research Council
802 (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013)
803 Consolidator Grant (FP7-THESEUS-617844), European Commission ITN (FP7-PloidyNet
804 607722), an ERC Advanced Grant (PROTEUS) from the European Research Council under
805 the European Union's Horizon 2020 research and innovation programme (835297) and
806 Chromavision from the European Union's Horizon 2020 research and innovation programme
807 (665233).
808

The TRACERx study (Clinicaltrials.gov no: NCT01888601) is sponsored by University College London (UCL/12/0279) and has been approved by an independent Research Ethics Committee (13/LO/1546). TRACERx is funded by Cancer Research UK (C11496/A17786) and coordinated through the Cancer Research UK and UCL Cancer Trials Centre.

The results published here are based in part on data generated by The Cancer Genome Atlas pilot project established by the NCI and the National Human Genome Research Institute. The data were retrieved through the database of Genotypes and Phenotypes (dbGaP) authorization (accession number phs000178.v9.p8). Information about TCGA and the constituent investigators and institutions of the TCGA research network can be found at http://cancergenome.nih.gov/. This project was enabled through access to the MRC eMedLab Medical Bioinformatics infrastructure, supported by the Medical Research Council (MR/L016311/1). In particular, we acknowledge the support of the High-Performance Computing at the Francis Crick Institute as well as the UCL Department of Computer Science Cluster and the support team.

The TRACERx Consortium

Charles Swanton (2,4,8), Mariam Jamal-Hanjani (2,8,9), Nicholas McGranahan (1,2), Carlos Martínez-Ruiz (1,2), Robert Bentham (1,2), Kevin Litchfield (2,3), Emilia L Lim (2,4), Crispin T Hiley (2,4), David A Moore (2,7,8), Thomas B K Watkins (4), Rachel Rosenthal (4), Maise Al Bakir (4), Roberto Salgado (5,6), Nicolai J Birkbak (11), Mickael Escudero (11), Aengus Stewart (11), Andrew Rowan (11), Jacki Goldman (11), Peter Van Loo (11), Richard Kevin Stone (11), Tamara Denner (11), Emma Nye (11), Sophia Ward (11), Stefan Boeing (11), Maria Greco (11), Jerome Nicod (11), Clare Puttick (11), Katey Enfield (11), Emma Colliver (11), Brittany Campbell (11), Alexander M Frankell (11), Daniel Cook (11), Mihaela Angelova (11), Alastair Magness (11), Chris Bailey (11), Antonia Toncheva (11), Krijn Dijkstra (11), Judit Kisistok (11), Mateo Sokac (11), Oriol Pich (11), Jonas Demeulemeester (11), Elizabeth Larose Cadieux (11), Carla Castignani (11), Krupa Thakkar (11), Hongchang Fu (11),

Takahiro Karasaki (11,12), Othman Al-Sawaf (11,12), Mark S Hill (11,40), Christopher Abbosh (12), Yin Wu (12), Selvaraju Veeriah (12), Robert E Hynds (12), Andrew Georgiou (12), Mariana Werner Sunderland (12), James L Reading (12), Sergio A Quezada (12), Karl S Peggs (12), Teresa Marafioti (12), John A Hartley (12), Helen L Lowe (12), Leah Ensell (12), Victoria Spanswick (12), Angeliki Karamani (12), Dhruva Biswas (12), Stephan Beck (12), Olga Chervova (12), Miljana Tanic (12), Ariana Huebner (12), Michelle Dietzen (12), James RM Black (12), Cristina Naceur-Lombardelli (12), Mita Afroza Akther (12), Haoran Zhai (12), Nnennaya Kanu (12), Simranpreet Summan (12), Francisco Gimeno-Valiente (12), Kezhong Chen (12), Elizabeth Manzano (12), Supreet Kaur Bola (12), Ehsan Ghorani (12), Marc Robert de Massy (12), Elena Hoxha (12), Emine Hatipoglu (12), Benny Chain (12), David R Pearce (12), Javier Herrero (12), Simone Zaccaria (12), Jason Lester (13), Fiona Morgan (14), Malgorzata Kornaszewska (14), Richard Attanoos (14), Haydn Adams (14), Helen Davies (14), Jacqui A Shaw (15), Joan Riley (15), Lindsay Primrose (15), Dean Fennell (15,16), Apostolos Nakas (16), Sridhar Rathinam (16), Rachel Plummer (16), Rebecca Boyles (16), Mohamad Tufail (16), Amrita Bajaj (16), Jan Brozik (16), Keng Ang (16), Mohammed Fiyaz Chowdhry (16), William Monteiro (17), Hilary Marshall (17), Alan Dawson (18), Sara Busacca (18), Domenic Marrone (18), Claire Smith (18), Girija Anand (19), Sajid Khan (19), Gillian Price (20), Mohammed Khalil (20), Keith Kerr (20), Shirley Richardson (20), Heather Cheyne (20), Joy Miller (20), Keith Buchan (20), Mahendran Chetty (20), Sylvie Dubois-Marshall (20), Sara Lock (21), Kayleigh Gilbert (21), Babu Naidu (22), Gerald Langman (22), Hollie Bancroft (22), Salma Kadiri (22), Gary Middleton (22), Madava Djearaman (22), Aya Osman (22), Helen Shackleford (22), Akshay Patel (22), Angela Leek (23), Nicola Totten (23), Jack Davies Hodgkinson (23), Jane Rogan (23), Katrina Moore (23), Rachael Waddington (23), Jane Rogan (23), Yvonne Summers (24), Raffaele Califano (24), Rajesh Shah (24), Piotr Krysiak (24), Kendadai Rammohan (24), Eustace Fontaine (24), Richard Booton (24), Matthew Evison (24), Stuart Moss (24), Juliette Novasio (24), Leena Joseph (24), Paul Bishop (24), Anshuman Chaturvedi (24), Helen Doran (24), Felice Granato (24), Vijay Joshi (24), Elaine Smith (24), Angeles Montero (24), Philip Crosbie (24,25,26), Fiona Blackhall (26,27), Lynsey Priest

(26,27), Matthew G Krebs (26,27), Caroline Dive (26,28), Dominic G Rothwell (26,28), Alastair Kerr (26,28), Elaine Kilgour (26,28), Katie Baker (27), Mathew Carter (27), Colin R Lindsay (27), Fabio Gomes (27), Jonathan Tugwood (28), Jackie Pierce (28), Alexandra Clipson (28), Roland Schwarz (29,30), Tom L Kaufmann (31,32), Matthew Huska (33), Zoltan Szallasi (34), Istvan Csabai (35), Miklos Diossy (35), Hugo Aerts (36,37), Charles Fekete (37), Gary Royle (38), Catarina Veiga (38), Marcin Skrzypski (39), David Lawrence (40), Martin Hayward (40), Nikolaos Panagiotopoulos (40), Robert George (40), Davide Patrini (40), Mary Falzon (40), Elaine Borg (40), Reena Khiroya (40), Asia Ahmed (40), Magali Taylor (40), Junaid Choudhary (40), Sam M Janes (40), Martin Forster (40), Tanya Ahmad (40), Siow Ming Lee (40), Neal Navani (40), Dionysis Papadatos-Pastos (40), Marco Scarci (40), Pat Gorman (40), Elisa Bertoja (40), Robert CM Stephens (40), Emilie Martinoni Hoogenboom (40), James W Holding (40), Steve Bandula (40), Ricky Thakrar (40), Radhi Anand (40), Kayalvizhi Selvaraju (40), James Wilson (40), Sonya Hessey (40), Paul Ashford (40), Mansi Shah (40), Marcos Vasquez Duran (40), Mairead MacKenzie (41), Maggie Wilcox (41), Allan Hackshaw (42), Yenting Ngai (42), Abigail Sharp (42), Cristina Rodrigues (42), Oliver Pressey (42), Sean Smith (42), Nicole Gower (42), Harjot Kaur Dhanda (42), Kitty Chan (42), Sonal Chakraborty (42), Christian Ottensmeier (43), Serena Chee (43), Benjamin Johnson (43), Aiman Alzetani (43), Judith Cave (43), Lydia Scarlett (43), Emily Shaw (43), Eric Lim (44), Paulo De Sousa (44), Simon Jordan (44), Alexandra Rice (44), Hilgardt Raubenheimer (44), Harshil Bhayani (44), Morag Hamilton (44), Lyn Ambrose (44), Anand Devaraj (44), Hema Chavan (44), Sofina Begum (44), Silviu I Buderi (44), Daniel Kaniu (44), Mpho Malima (44), Sarah Booth (44), Andrew G Nicholson (44), Nadia Fernandes (44), Christopher Deeley (44), Pratibha Shah (44), Chiara Proli (44), Kelvin Lau (45), Michael Sheaff (45), Peter Schmid (45), Louise Lim (45), John Conibear (45), Madeleine Hewish (46), Sarah Danson (47), Jonathan Bury (47), John Edwards (47), Jennifer Hill (47), Sue Matthews (47), Yota Kitsanta (47), Jagan Rao (47), Sara Tenconi (47), Laura Socci (47), Kim Suvarna (47), Faith Kibutu (47), Patricia Fisher (47), Robin Young (47), Joann Barker (47), Fiona Taylor (47), Kirsty Lloyd (47), Michael Shackcloth (48), Julius Asante-Siaw (48), John Gosney (49), Teresa Light (50), Tracey Horey (50), Peter

Russell (50), Dionysis Papadatos-Pastos (50), Kevin G Blyth (51), Craig Dick (51), Andrew Kidd (51), Alan Kirk (52), Mo Asif (52), John Butler (52), Rocco Bilancia (52), Nikos Kostoulas (52), Mathew Thomas (52), Gareth A Wilson (53)

(11) The Francis Crick Institute, London, United Kingdom

(12) University College London Cancer Institute, London, United Kingdom

(13) Swansea Bay University Health Board, Swansea, United Kingdom

(14) Cardiff & Vale University Health Board, Cardiff, United Kingdom

(15) Cancer Research Centre, University of Leicester, Leicester, United Kingdom

(16) Leicester University Hospitals, Leicester, United Kingdom

(17) National Institute for Health Research Leicester Respiratory Biomedical Research Unit, Leicester, United Kingdom

(18) University of Leicester, Leicester, United Kingdom

(19) Barnet & Chase Farm Hospitals, Barnet, United Kingdom

(20) Aberdeen Royal Infirmary, Aberdeen, United Kingdom

(21) The Whittington Hospital NHS Trust, London, United Kingdom

(22) University Hospital Birmingham NHS Foundation Trust, Birmingham, United Kingdom

(23) Manchester Cancer Research Centre Biobank, Manchester, United Kingdom

(24) Wythenshawe Hospital, Manchester University NHS Foundation Trust, Manchester, United Kingdom

(25) Division of Infection, Immunity and Respiratory Medicine, University of Manchester, Manchester, UK

(26) Cancer Research UK Lung Cancer Centre of Excellence, University of Manchester, Manchester, UK

(27) Christie NHS Foundation Trust, Manchester, United Kingdom

(28) Cancer Research UK Manchester Institute, University of Manchester, Manchester, UK

(29) Berlin Institute for Medical Systems Biology, Max Delbrueck Center for Molecular Medicine, Berlin, Germany

(30) German Cancer Consortium (DKTK), partner site Berlin, Berlin, Germany

(31) Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Robert-Rössle-Str 10, 13125, Berlin, Germany

(32) BIFOLD, Berlin Institute for the Foundations of Learning and Data, Berlin, Germany

(33) Berlin Institute for Medical Systems Biology, Max Delbrueck Center for Molecular Medicine, Berlin, Germany

(34) Danish Cancer Society Research Center, Copenhagen, Denmark

(35) Department of Physics of Complex Systems, ELTE Eötvös Loránd University, Budapest, Hungary

(36) Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical School, Boston, Massachusetts, USA

(37) Radiology and Nuclear Medicine, CARIM & GROW, Maastricht University, Maastricht, The Netherlands

(38) Department of Medical Physics and Bioengineering, University College London Cancer Institute, London, United Kingdom

(39) Department of Oncology and Radiotherapy, Medical University of Gdańsk, ul Dębinki 7, 80-211, Gdańsk, Poland

(40) University College London Hospitals, London, United Kingdom

(41) Independent Cancer Patients Voice, London, United Kingdom

(42) Cancer Research UK & UCL Cancer Trials Centre, London, United Kingdom

(43) University Hospital Southampton NHS Foundation Trust, Southampton, United Kingdom

(44) Royal Brompton and Harefield NHS Foundation Trust, London, United Kingdom

(45) Barts Health NHS Trust, London, United Kingdom

(46) Ashford and St Peter's Hospitals NHS Foundation Trust, Chertsey, United Kingdom

(47) Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, United Kingdom

(48) Liverpool Heart and Chest Hospital NHS Foundation Trust, Liverpool, United Kingdom

(49) Royal Liverpool University Hospital, Liverpool, United Kingdom

(50) The Princess Alexandra Hospital NHS Trust, Harlow, United Kingdom

(51) NHS Greater Glasgow and Clyde, Glasgow, United Kingdom

(52) Golden Jubilee National Hospital, Clydebank, United Kingdom

(53) Achilles Therapeutics UK Limited, London, United Kingdom

References

1. Rosenthal, R. *et al.* Neoantigen-directed immune escape in lung cancer evolution. *Nature* **567**, 479–485 (2019).
2. Litchfield, K. *et al.* Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell* **184**, 596-614.e14 (2021).
3. Robert, C. *et al.* Ipilimumab plus Dacarbazine for Previously Untreated Metastatic Melanoma. *N. Engl. J. Med.* **364**, 2517–2526 (2011).
4. Schadendorf, D. *et al.* Pooled analysis of long-term survival data from phase II and phase III trials of ipilimumab in unresectable or metastatic melanoma. *J. Clin. Oncol.* **33**, 1889–1894 (2015).
5. Goodman, A. M. *et al.* Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Mol. Cancer Ther.* **16**, 2598–2608 (2017).
6. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 16910–16915 (2010).
7. Favero, F. *et al.* Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).
8. Shen, R. & Seshan, V. FACETS : Fraction and Allele-Specific Copy Number Estimates from Tumor Sequencing. *Meml. Sloan-Kettering Cancer Center, Dept. Epidemiol. Biostat. Work. Pap. Ser.* **1**, 50 (2015).
9. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).

10. López, S. *et al.* Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. *Nat. Genet.* **52**, 283–293 (2020).
11. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
12. Levy, E. *et al.* Immune DNA signature of T-cell infiltration in breast tumor exomes. *Sci. Rep.* **6**, 1–10 (2016).
13. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
14. Danaher, P. *et al.* Pan-cancer adaptive immune resistance as defined by the Tumor Inflammation Signature (TIS): Results from The Cancer Genome Atlas (TCGA). *J. Immunother. Cancer* **6**, 1–17 (2018).
15. Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science (80-.).* **355**, (2017).
16. Aran, D., Hu, Z. & Butte, A. J. xCell: Digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 1–14 (2017).
17. Li, T. *et al.* TIMER: A web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer Res.* **77**, e108–e110 (2017).
18. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
19. Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *Elife* **6**, 1–25 (2017).
20. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
21. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
22. Yokoyama, A. *et al.* Age-related remodelling of oesophageal epithelia by

mutated cancer drivers. *Nature* **565**, 312–317 (2019).

23. Poore, G. D. *et al.* Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567–574 (2020).

24. Watkins, T. B. K. *et al.* Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature* **587**, 126–132 (2020).

25. Jongsma, M. L. M. *et al.* The SPPL3-Defined Glycosphingolipid Repertoire Orchestrates HLA Class I-Mediated Immune Responses. *Immunity* **54**, 132-150.e9 (2021).

26. AbdulJabbar, K. *et al.* Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. *Nat. Med.* **26**, 1054–1062 (2020).

27. Schwartz, L. H. *et al.* RECIST 1.1 - Update and clarification: From the RECIST committee. *Eur. J. Cancer* **62**, 132–137 (2016).

28. Shim, J. H. *et al.* HLA-corrected tumor mutation burden and homologous recombination deficiency for the prediction of response to PD-(L)1 blockade in advanced non-small-cell lung cancer patients. *Ann. Oncol.* **31**, 902–911 (2020).

29. Rizvi, N. A. *et al.* Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science (80-.).* **348**, 124–128 (2015).

30. Conforti, F. *et al.* Sex-based dimorphism of anticancer immune response and molecular mechanisms of immune evasion. *Clin. Cancer Res.* clincanres.0136.2021 (2021). doi:10.1158/1078-0432.ccr-21-0136

31. Capone, I., Marchetti, P., Ascierto, P. A., Malorni, W. & Gabriele, L. Sexual Dimorphism Of Immune Responses: A new perspective in cancer immunotherapy. *Front. Immunol.* **9**, 1–8 (2018).

32. van der Spek, J., Groenwold, R. H. H., van der Burg, M. & van Montfrans, J. M. TREC Based Newborn Screening for Severe Combined Immunodeficiency Disease: A Systematic Review. *J. Clin. Immunol.* **35**, 416–430 (2015).

Additional References

33. Kuchenbecker, L. *et al.* IMSEQ-A fast and error aware approach to immunogenetic sequence analysis. *Bioinformatics* **31**, 2963–2971 (2015).

34. Wood, D. E. & Salzberg, S. L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, (2014).

35. Middleton, G. *et al.* The National Lung Matrix Trial of personalized therapy in lung cancer. *Nature* **583**, 807–812 (2020).

36. Wang, K. *et al.* PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).

37. Brastianos, P. K. *et al.* Genomic characterization of brain metastases reveals branched evolution and potential therapeutic targets. *Cancer Discov.* **5**, 1164–1177 (2015).

38. Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46**, 225–233 (2014).

39. Gerlinger, M. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).

40. Harbst, K. *et al.* Multiregion whole-exome sequencing uncovers the genetic evolution and mutational heterogeneity of early-stage metastatic melanoma. *Cancer Res.* **76**, 4765–4774 (2016).

41. Lamy, P. *et al.* Paired exome analysis reveals clonal evolution and potential therapeutic targets in urothelial carcinoma. *Cancer Res.* **76**, 5894–5906 (2016).

42. Savas, P. *et al.* The Subclonal Architecture of Metastatic Breast Cancer: Results from a Prospective Community-Based Rapid Autopsy Program “CASCADE”. *PLoS Med.* **13**, 1–25 (2016).

43. Suzuki, H. *et al.* Mutational landscape and clonal architecture in grade II and III gliomas. *Nat. Genet.* **47**, 458–468 (2015).

44. Turajlic, S. *et al.* Deterministic Evolutionary Trajectories Influence Primary Tumor

Growth: TRACERx Renal. *Cell* **173**, 595-610.e11 (2018).

45. Messaoudene, M. *et al.* T-cell bispecific antibodies in node-positive breast cancer: Novel therapeutic avenue for MHC class i loss variants. *Ann. Oncol.* **30**, 934–944 (2019).

46. Snyder, A. *et al.* Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. *N. Engl. J. Med.* **371**, 2189–2199 (2014).

47. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science (80-.).* **352**, 207–212 (2016).

48. Hugo, W. *et al.* Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* **165**, 35–44 (2016).

49. Riaz, N. *et al.* Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell* **171**, 934-949.e15 (2017).

50. Cristescu, R. *et al.* Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. *Science (80-.).* **362**, (2018).

51. Snyder, A. *et al.* Contribution of systemic and somatic factors to clinical response and resistance to PD-L1 blockade in urothelial cancer: An exploratory multi-omic analysis. *PLoS Med.* **14**, 1–24 (2017).

52. Mariathasan, S. *et al.* TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature* **554**, 544–548 (2018).

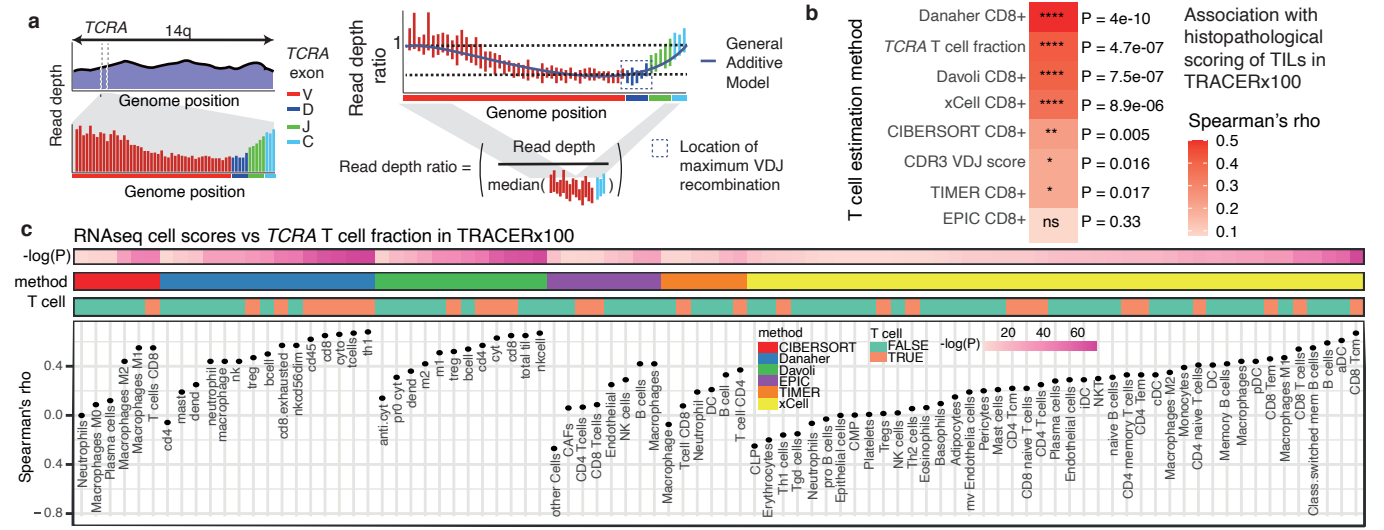
53. McDermott, D. F. *et al.* Clinical activity and molecular correlates of response to atezolizumab alone or in combination with bevacizumab versus sunitinib in renal cell carcinoma. *Nat. Med.* **24**, 749–757 (2018).

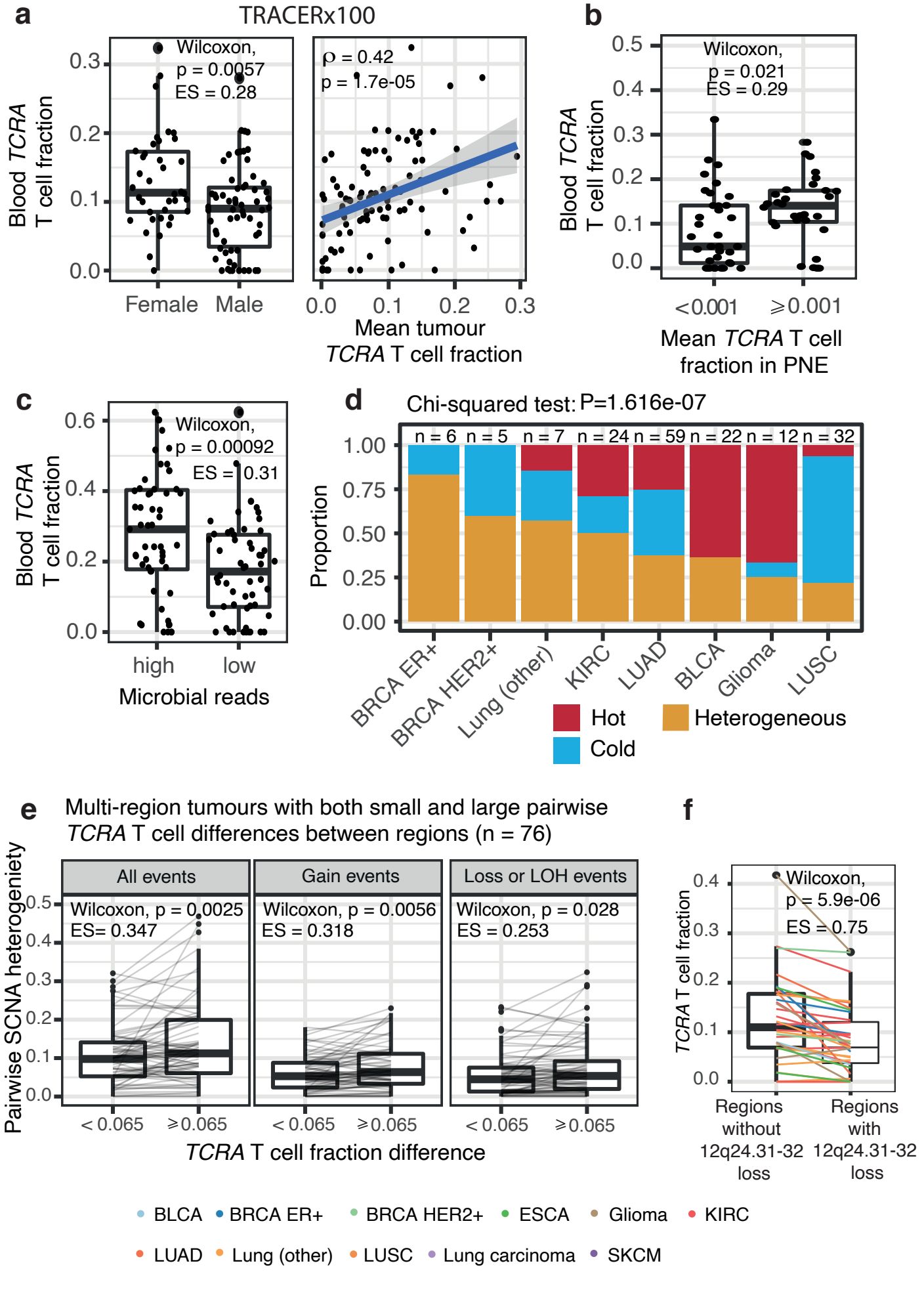
54. Le, D. T. *et al.* PD-1 blockade in tumors with mismatch repair deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).

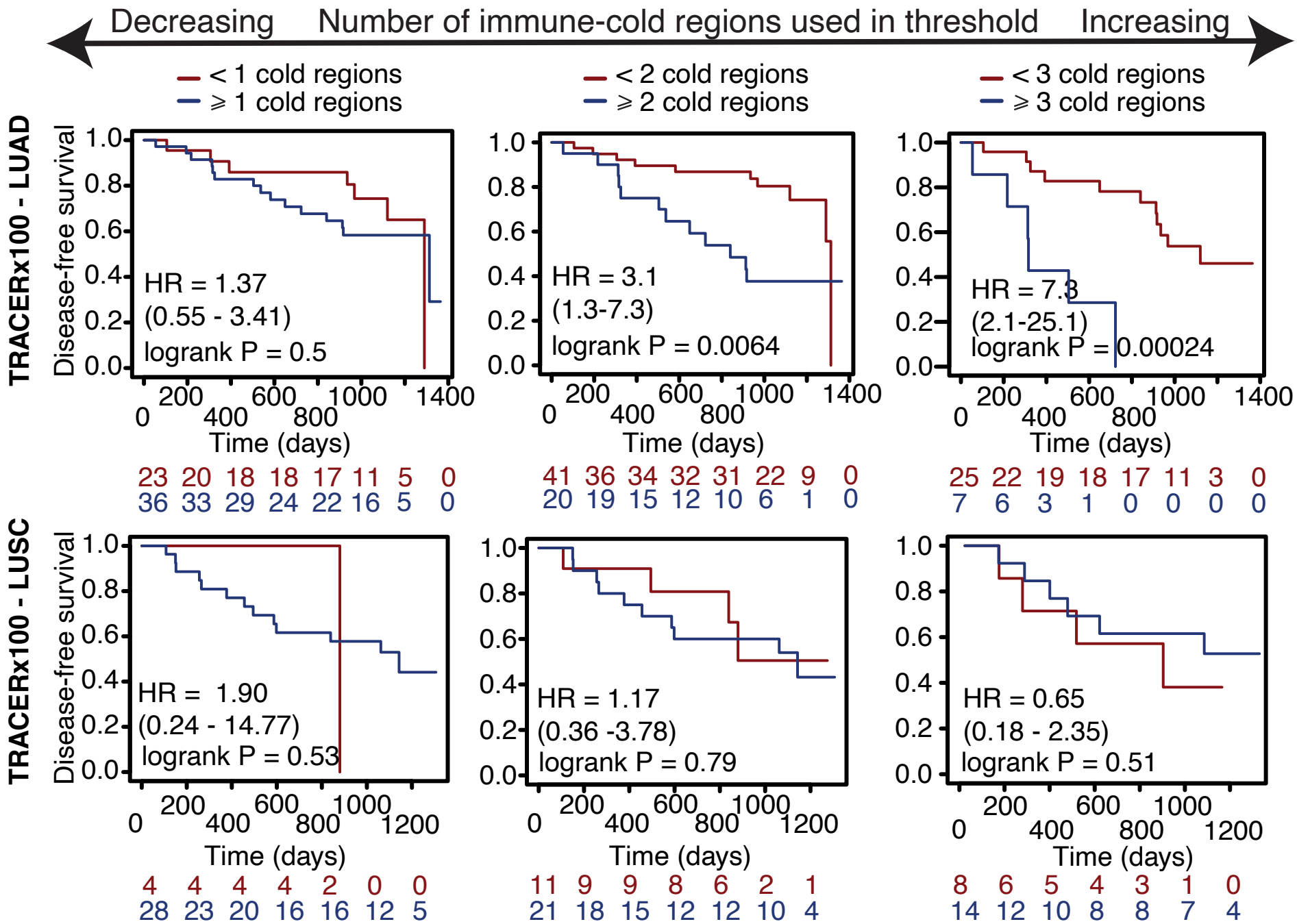
55. Hendry, S. *et al. Assessing Tumor-infiltrating Lymphocytes in Solid Tumors*. *Advances In Anatomic Pathology* **24**, (2017).

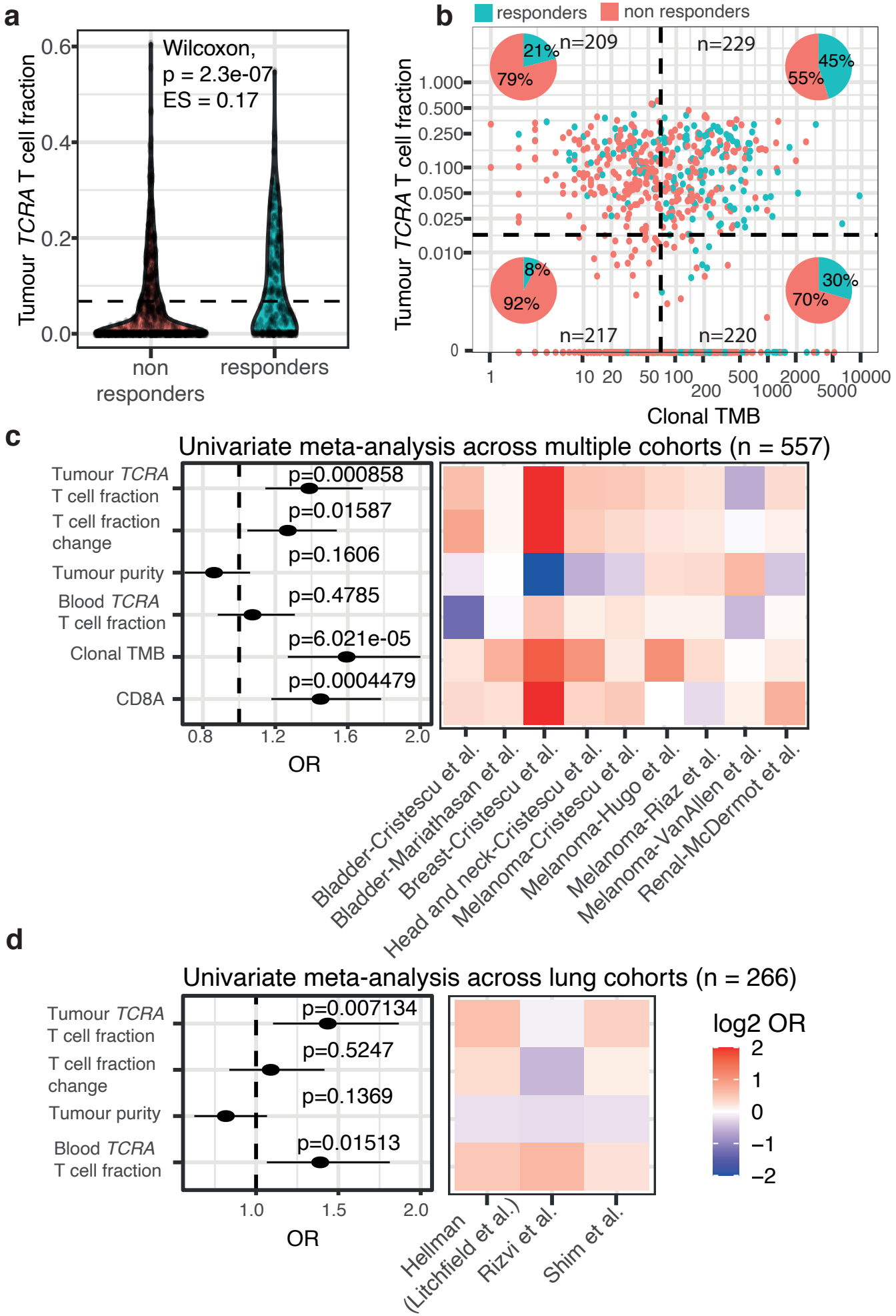
56. Denkert, C. *et al.* Standardized evaluation of tumor-infiltrating lymphocytes in breast cancer: Results of the ring studies of the international immuno-oncology

biomarker working group. *Mod. Pathol.* **29**, 1155–1164 (2016).


a
TCRA
14q
Read depth
Genome position
TCRA exon
V
D
J
C
Read depth ratio
1
General Additive Model
Location of maximum VDJ recombination
Read depth ratio = Read depth / median(
b
T cell estimation method
Danaher CD8+
TCRA T cell fraction
Davoli CD8+
xCell CD8+
CIBERSORT CD8+
CDR3 VDJ score
TIMER CD8+
EPIC CD8+
P = 4e-10
P = 4.7e-07
P = 7.5e-07
P = 8.9e-06
P = 0.005
P = 0.016
P = 0.017
P = 0.33
Association with histopathological scoring of TILs in TRACERx100
Spearman's rho
0.5
0.4
0.3
0.2
0.1
c
RNAseq cell scores vs TCRA T cell fraction in TRACERx100
-log(P)
method
T cell
Spearman's rho
method
CIBERSORT
Danaher
Davoli
EPIC
TIMER
xCell
T cell
FALSE
TRUE
-log(P)
20
40
60
Neutrophils
Macrophages M0
Plasma cells
Macrophages M2
Macrophages M1
T cells CD8
cd4
mast
dend
neutrophil
macrophage
nk
treg
bcell
cd8.exhausted
nkcd56dim
cd45
cd8
cyto
tcells
th1
anti.cyt
pr0 cyt
dend
m2
m1
treg
bcell
cd4
cyt
cd8
total til
nkcell
other Cells
CAFs
CD4 Tcells
CD8 Tcells
Endothelial
NK cells
B cells
Macrophages
Macrophage
Tcell CD8
Neutrophil
DC
B cell
T cell CD4
CLP
Erythrocytes
Th1 cells
Tgd cells
Neutrophils
pro B cells
Epithelial cells
CMP
Platelets
Tregs
NK cells
Th2 cells
Eosinophils
Basophils
Adipocytes
mv Endothelia cells
Pericytes
Mast cells
CD4 Tcm
CD8 naive T cells
CD4 T cells
Plasma cells
Endothelial cells
iDC
NKT
naive B cells
CD4 memory T cells
CD4 Tem
cDC
Macrophages M2
Monocytes
CD4 naive T cells
DC
Memory B cells
Macrophages M1
pDC
CD8 Tem
CD8 T cells
Class.switched mem B cells
B cells
aDC
CD8 Tcm


a
TRACERx100
Blood TCRA T cell fraction
Wilcoxon, p = 0.0057 ES = 0.28
Female
Male
ρ = 0.42 p = 1.7e-05
Mean tumour TCRA T cell fraction
b
Wilcoxon, p = 0.021 ES = 0.29
<0.001
⩾0.001
Mean TCRA T cell fraction in PNE
c
Wilcoxon, p = 0.00092 ES = 0.31
high
low
Microbial reads
d
Chi-squared test: P=1.616e-07
Proportion
n = 6 n = 5 n = 7 n = 24 n = 59 n = 22 n = 12 n = 32
BRCA ER+
BRCA HER2+
Lung (other)
KIRC
LUAD
BLCA
Glioma
LUSC
Hot
Cold
Heterogeneous
e
Multi-region tumours with both small and large pairwise TCRA T cell differences between regions (n = 76)
All events
Wilcoxon, p = 0.0025 ES= 0.347
Gain events
Wilcoxon, p = 0.0056 ES = 0.318
Loss or LOH events
Wilcoxon, p = 0.028 ES = 0.253
Pairwise SCNA heterogeneity
< 0.065
⩾0.065
TCRA T cell fraction difference
f
TCRA T cell fraction
Wilcoxon, p = 5.9e-06 ES = 0.75
Regions without 12q24.31-32 loss
Regions with 12q24.31-32 loss
BLCA
BRCA ER+
BRCA HER2+
ESCA
Glioma
KIRC
LUAD
Lung (other)
LUSC
Lung carcinoma
SKCM


Decreasing
Number of immune-cold regions used in threshold
Increasing
TRACERx100 - LUAD
TRACERx100 - LUSC
Disease-free survival
Time (days)
< 1 cold regions
⩾ 1 cold regions
HR = 1.37
(0.55 - 3.41)
logrank P = 0.5
23 20 18 18 17 11 5 0
36 33 29 24 22 16 5 0
< 2 cold regions
⩾ 2 cold regions
HR = 3.1
(1.3-7.3)
logrank P = 0.0064
41 36 34 32 31 22 9 0
20 19 15 12 10 6 1 0
< 3 cold regions
⩾ 3 cold regions
HR = 7.3
(2.1-25.1)
logrank P = 0.00024
25 22 19 18 17 11 3 0
7 6 3 1 0 0 0 0
HR = 1.90
(0.24 - 14.77)
logrank P = 0.53
4 4 4 4 2 0 0
28 23 20 16 16 12 5
HR = 1.17
(0.36 -3.78)
logrank P = 0.79
11 9 9 8 6 2 1
21 18 15 12 12 10 4
HR = 0.65
(0.18 - 2.35)
logrank P = 0.51
8 6 5 4 3 1 0
14 12 10 8 8 7 4


a
Tumour TCRA T cell fraction
0.6
0.4
0.2
0.0
Wilcoxon,
p = 2.3e-07
ES = 0.17
non responders
responders
b
responders
non responders
Tumour TCRA T cell fraction
1.000
0.500
0.250
0.100
0.050
0.025
0.010
0
n=209
n=229
n=217
n=220
21%
79%
45%
55%
8%
92%
30%
70%
1
10
20
50
100
200
500
1000
2000
5000
10000
Clonal TMB
c
Univariate meta-analysis across multiple cohorts (n = 557)
Tumour TCRA T cell fraction
T cell fraction change
Tumour purity
Blood TCRA T cell fraction
Clonal TMB
CD8A
p=0.000858
p=0.01587
p=0.1606
p=0.4785
p=6.021e-05
p=0.0004479
0.8
1.2
1.6
2.0
OR
Bladder-Cristescu et al.
Bladder-Mariathasan et al.
Breast-Cristescu et al.
Head and neck-Cristescu et al.
Melanoma-Cristescu et al.
Melanoma-Hugo et al.
Melanoma-Riaz et al.
Melanoma-VanAllen et al.
Renal-McDermot et al.
d
Univariate meta-analysis across lung cohorts (n = 266)
Tumour TCRA T cell fraction
T cell fraction change
Tumour purity
Blood TCRA T cell fraction
p=0.007134
p=0.5247
p=0.1369
p=0.01513
1.0
1.5
2.0
OR
Hellman (Litchfield et al.)
Rizvi et al.
Shim et al.
log2 OR
2
1
0
−1
−2